# Chapter 15
# Generalised Additive Models

**Robert M. West**

## 15.1 Introduction

The inclusion of continuous covariates in generalised linear models is common in epidemiological applications. For example, age and deprivation are very common confounders and so are often 'adjusted for'. Sometimes, although covariates are continuous, they are entered in discretised form. This is one method employed to account for nonlinearity and is discussed in more detail below. The issue concerning this chapter is that covariates need not enter a generalised linear model merely as linear terms.

Specifically consider the outcome variable to be mortality and that a logistic regression is used to model the effects of covariates. The model will be used to explain mortality rather than simply to predict mortality. Epidemiological study focuses on an exposure, which enters as a covariate. Age often has a clinically and statistically large impact on mortality and, although often just a nuisance variable, needs to be included as a covariate. It is sometimes, but not often, plausible that the log odds of mortality increases linearly with age. More commonly the relationship has greater complexity. If age is poorly modelled then the estimate of the exposure of interest will be less accurate and biased. The log odds of mortality and age is most likely to increase with age and is plausibly smooth: the question becomes how the best model fit is obtained. The answer might be to use a transformation of the covariate, consider higher-order terms, to fit splines, or to make use of the techniques employed in Generalised Additive Models (GAMs).

Nonlinearity is not the only consideration that motivates the use of GAMs. The difficulty of interactions, see Chap. 16, for continuous covariates has been noted. In particular determining the functional form of second- and higher-order

R.M. West (✉)
Division of Biostatistics, Centre for Epidemiology and Biostatistics, Leeds Institute of Genetics, Health & Therapeutics, University of Leeds, Leeds, UK
e-mail: r.m.west@leeds.ac.uk

Y.-K. Tu and D.C. Greenwood (eds.), *Modern Methods for Epidemiology*,
DOI 10.1007/978-94-007-3024-3_15, 

interactions is even more challenging than for a single main effect. The techniques available with GAMs provide a suitable means to tackle this ferocious challenge.

The first step however is to focus on the challenges of nonlinearity for a single main effect. Throughout this chapter data from a study of sympathetic nerve activity has been selected to illustrate issues and procedures.

## 15.2 Sympathetic Nerve Activity: Basic Model

Sympathetic nerve activity (*sna*) is known to increase with *age* and so is a convenient example for the topic of this chapter. Further, there is a complex relationship with systolic blood pressure (*sbp*) as well, so that there are two continuous covariates to explore in models of *sna* (Burns et al. 2007). The setting for this example is a study where 172 volunteers were recruited in order to investigate certain aspects of the variation of *sna* between individuals. For simplicity here only the effects of *sex*, *age*, and *sbp* on *sna* will be considered, and although the causal relationship might be debated, in this and the next chapter, *sna* is taken to depend upon the other variables.

The outcome *sna* is a measurement on a continuous scale, *sex* is a dichotomous covariate (factor), and as mentioned above, *age* and *sbp* are continuous covariates. A basic model will fit just linear terms as covariates. All modelling will be undertaken in R since this statistical language is widely available (R Development Core Team 2010) and has good capabilities, once the relevant libraries have been downloaded. In R, models are specified by notation suggested by Wilkinson and Rogers (1973), and is straightforward to follow. The basic model is specified by `sna ~ as.factor(sex) + age + sbp` and the fitted model yields the results given in Table 15.1.

Note that for this model the adjusted sum of squares is 0.60: 60% of variation is explained by the model. The errors were also explored through plots, and it was seen that the residual plot against the fitted values, the normal QQ plot, the scale location plot and the leverage plots were all satisfactory. This is also true for all the subsequent residual plots in this chapter.

For this basic model, the effects of *age* and *sbp* are clear: they are simply linear terms. For completeness, and to permit comparison with later plots, graphical representations are provided in Figs. 15.1 and 15.2. These include rug plots along the abscissas to indicate for which ages and SBPs measurements of *sna* have been recorded. Note also the ranges of the ordinates.

**Table 15.1** Table of coefficients for the basic model

| Coefficient | Estimate | 95% CI | *p*-value |
|---|---|---|---|
| Intercept | −16.6 | (−26.5, −6.6) | 0.00123 |
| Male | 7.3 | (3.8, 10.8) | 5.37e-05 |
| Age | 0.70 | (3.8, 10.8) | 6.24e-12 |
| SBP | 0.25 | (0.16, 0.33) | 4.54e-08 |

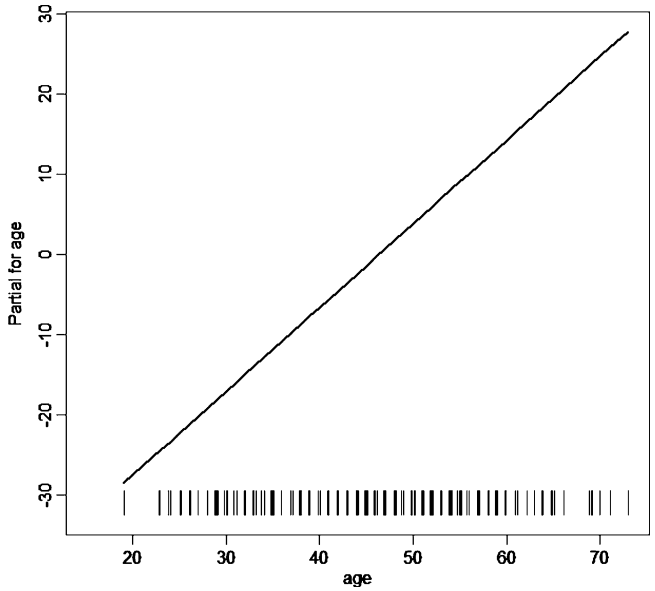


**Fig. 15.1** Term plot for linear *age*

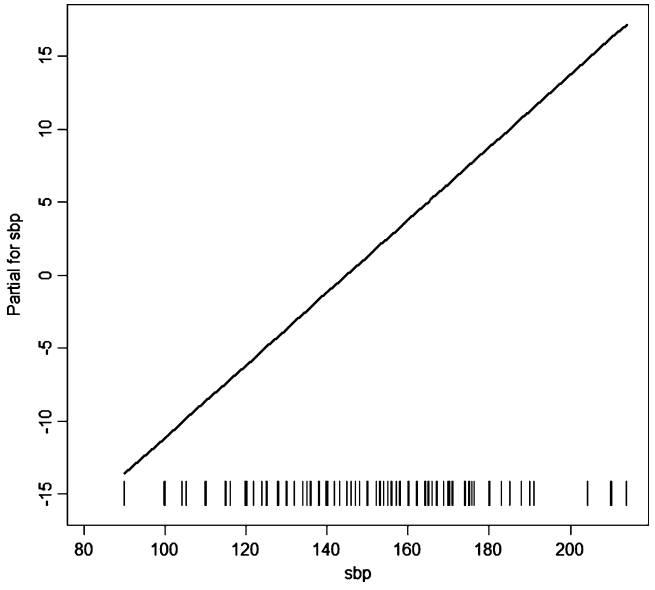


**Fig. 15.2** Term plot for linear SBP

The use of discretisation of age into age groups has a long history. John Graunt (1662) is one of the earliest to publish material (life tables) and establish this methodology that has been exploited to great effect by modern insurance companies as just one example. As more data is available, the width of the age groups can be diminished, any errors due to discretisation will be minimal, and age can be considered to be modelled sufficiently well.

Such fine discretisation is not however undertaken throughout epidemiology, even when sample sizes are large: widths of age groups of 5 or 10 years can be found. There is an issue of parsimony in the model. If *m* age groups are to be used then (*m-1*) variables are required to model age. Then a polynomial of degree (*m-1*) is just as parsimonious and should be considered, see Sect. 15.4.

Discretisation might be favoured for reasons of interpretation, especially with logistic regression. For example age might be discretised as: Under 60, 60–69, 70–85, and Over 85 years. Then logistic regression delivers three odds ratios comparing the odds of mortality for persons in the three older groups with those in the youngest group. Interpretation is very simple in relation to the age effect. Effectively though, age has been modelled as a step function. An individual of age 69 steps up their risk on their seventieth birthday. There is certainly a discretisation error. The main concern though is that inaccuracy in modelling age will result in inaccuracy and bias for the role of other covariates including the exposure of interest.

Another concern about discretisation is that the number of groups and the group boundaries need to be chosen. There may be clinical or political reasons for specifying boundaries, such as achieving adult status at age 18, achieving retirement age at 65, etc. The results achieved for all covariate coefficients will differ when boundaries are changed. From a modelling perspective, the boundaries may be chosen for example by minimising the Akaike Information Criterion (AIC), although this may lead to what seem strange boundary values that once more lead to interpretation difficulties, albeit a fascinating challenge to obtain an interpretation.

Where there is a choice of the number of groups and their boundaries, there is 'temptation' to choose them to deliver the coefficient values of other covariates that are most favoured—especially if the main exposure has a coefficient close to statistical significance, but these issues are always present in complex modelling situations.

When a continuous variable is discretised, it is easy to define a further category of 'missing' when values are not recorded for some participants. This has great appeal if such an approach is appropriate for the modelling of missing values—for example where values are missing at random. In other circumstances however this could be disastrous. Consider the case where age is withheld by either the very young or very old for reasons of identifiability of those with a rare disease. Then such a category is misleading and it might be more appropriate to consider an imputation technique to handle missingness.

It is possible that some continuous covariates are discretised due to doubt about their true nature. An example might be a score from a psychometric test, which is not truly continuous, being the sum of (weighted) responses to a questionnaire.

Established thresholds might be used, for example defining a patient as depressed if he/she scores over a certain value on a depression scale. The underlying scale may not regarded by all as ordinal, let alone continuous. It is not clear if the use of the established thresholds improves matters. Information is lost regarding a variable whenever it is discretised and so error is introduced into the model, and the model cannot be improved by adding discretisation error. In such circumstances the variable might be considered as measured with error: refer to Chap. 3.

Deprivation scores are also composites and their full validity is sometimes in doubt: specifically the deprivation score of an area is associated with an individual. Often in epidemiological studies continuous measures of deprivation such as the Index of Multiple Deprivation or Townsend Score, are divided into fifths. This allows for nonlinearity but again there must be discretisation error as well as measurement error. The discretisation error might be avoided by using higher-order terms of deprivation scores: polynomial expressions of deprivation. Plots of the impact of deprivation on the modelled outcome will be required to facilitate interpretation as seen below in Sect. 15.4.

## 15.3 Sympathetic Nerve Activity: Discretisation

The basic model provided a fit for all three covariates with highly significant values for the three coefficients, but nonlinearity does potentially exist and an investigation is warranted. Here both *age* and *sbp* are discretised into five categories forming the new variables *agegp* and *sbpgp*. Cut points for *age* were taken as 30, 40, 50, and 60 years. Those for *sbp* were taken as 120, 140, 160, 180 mm Hg. For both variables the categories are all reasonably evenly populated whilst the cut points are easy to interpret. Values of *sbp* above 140 mm Hg suggest hypertension and so 140 mm Hg has some clinical meaning.

The model is specified by `sna ~ as.factor(sex) + as.factor(agegp) + as.factor(sbpgp)`, and results given in Table 15.2.

The contributions of the covariates *age* and *sbp* are expressed graphically in Figs. 15.3 and 15.4. Note that one clear effect is that the ranges of the effects are much reduced from those in the basic model: compare the graphs. For this model with discretised covariates, the adjusted $R^2 = 0.63$, so that on the basis of the proportion of variation that is represented, the model with discretised *age* and *sbp* is preferred to the basic model.

From Table 15.2 it is strongly tempting to coalesce some categories, thus improving the adjusted $R^2$. In particular the exact match of the category boundary for *sbp* with the definition of hypertension *sbp* = 140 mm Hg, is extremely tempting. Such a data-driven approach however can be regarded as over-fitting to the dataset. If disctretisation is to be employed, it is advisable to fix the boundaries of all categories before fitting to the data.

**Table 15.2** Table of coefficients for the model with discretised covariates

| Coefficient | Estimate | 95% CI | *p*-value |
|---|---|---|---|
| Intercept | 25.0 | (20.1,29.9) | $<$ 2e-16 |
| Male | 7.3 | (4.0,10.7) | 2.93e-5 |
| Age [30,40) | 13.4 | (6.8,20.0) | 9.33e-5 |
| Age [40,50) | 20.3 | (3.9,26.7) | 2.99e-9 |
| Age [50,60) | 22.7 | (15.9,29.4) | 5.13e-10 |
| Age $\geq$ 60 | 28.8 | (20.5,37.1) | 1.51e-10 |
| SBP [120,140) | 2.2 | ($-$3.1,7.5) | 0.413 |
| SBP [140,160) | 17.9 | (12.1,23.7) | 7.29e-9 |
| SBP [160,180) | 16.0 | (10.3,21.7) | 1.22e-7 |
| SBP $\geq$ 180 | 15.0 | (6.8,23.2) | 0.000419 |

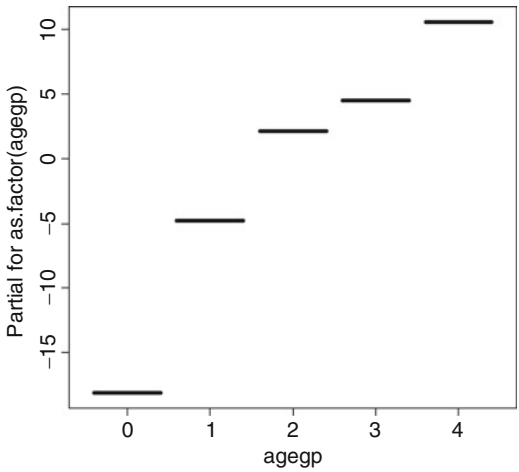


**Fig. 15.3** Term plot for discretised *age*

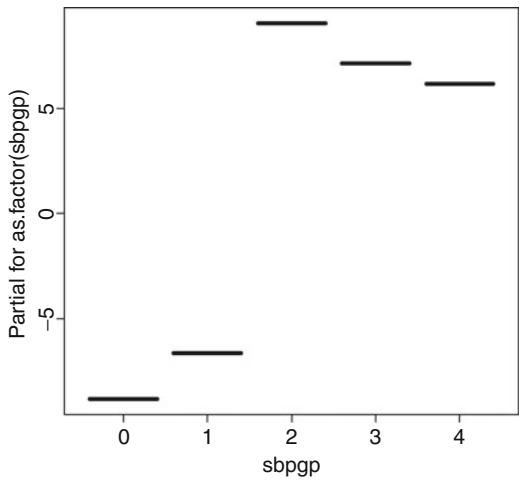


**Fig. 15.4** Term plot for discretised *sbp*

## 15.4 Higher-Order Terms

A very straightforward way to check if a covariate should appear as a linear term only is to fit higher-order terms and test their significance. Thus if *age* has been entered as a linear term, then adding $age^2$ and $age^3$ will reveal if the relationship is more complex. Interpreting the contribution of that covariate is not so easy to from a table of coefficients. It will be necessary to construct the fitted polynomial and plot it in order to make the position clear.

A phrase attributed to George Box is commonly cited: 'all models are wrong'. By better fitting of covariates, the models will be improved and the effects of exposures better assessed. Hence the added complexity of polynomial terms can be justified. Measures of fit of models can be used to balance complexity against fit such as adjusted $R^2$, AIC, and others. There are often automated searches available within software packages to obtain the best fit against these criteria. For example, R has a function `R::leaps::regsubsets()`, Miller (2002), that can search for the best fit by adjusted $R^2$, and the functions `R::stats::step()` and `R::MASS::stepAIC()` Venables and Ripley (2002) to search for the best fit by AIC. So there can be few excuses for not exploring this approach.

Searching higher-order terms can be made more efficient and robust by using orthogonal polynomials, Kennedy and Gentle (1980), due to increased numerical stability and the ease with which the best degree can be determined: orthogonality helps. In R, the function `R::stats::poly()` provides this ability. The function `R::stats::termplot()` can be used to display the functional representation and its influence on the outcome variable.

## 15.5 Sympathetic Nerve Activity: Higher-Order Terms

For the illustrative example, orthogonal polynomials were chosen, the formula in the R code being `sna ~ as.factor(sex) + poly(age,3) + poly(sbp,3)`.

From Table 15.3, the impact of the covariates on the outcome *sna* is not immediately clear. This is where graphical representations become important. Figures 15.5 and 15.6 demonstrate the effect of *age* and *sbp* effectively. Comparing the graphical figures for each of the models that have been fitted, it appears that the effect of *age* gives the largest range of effect in the model with higher-order terms, the youngest age resulting in a sizable decrease in *sna*: see Sect. 15.6 below for further comment.

Inspecting Fig. 15.6, the final downturn in the effect of SBP can be seen from the rug plot to be based on just a few measurements where *sbp* is above 200 mm Hg. Considering also the marginal statistical significance ($p = 0.0744$) of the cubic term for *sbp*, many might consider refitting with only a quadratic polynomial for *sbp*. The cubic representation is chosen here to identify that there is an issue of how best to identify the degree of polynomial representations of covariates in general: this issue is dealt with in Sect. 15.8.

**Table 15.3** Table of coefficients for model with higher-order terms of covariates

| Coefficient | Estimate | 95% CI | *p*-value |
|---|---|---|---|
| Intercept | 52.6 | (50.2,54.9) | < 2e-16 |
| Male | 6.3 | (2.9,9.6,) | 0.000265 |
| Poly(age,3) 1 | 108.9 | (81.4,136.4) | 5.95e-13 |
| Poly(age,3) 2 | −27.3 | (−49.9,−4.8) | 0.0179 |
| Poly(age,3) 3 | 42.2 | (20.7,63.7) | 0.000152 |
| Poly(sbp,3) 1 | 85.7 | (58.2,113.2) | 5.60e-9 |
| Poly(sbp,3) 2 | −24.0 | (−46.5,−1.5) | 0.0364 |
| Poly(sbp,3) 3 | −19.4 | (−40.7,1.9) | 0.0744 |

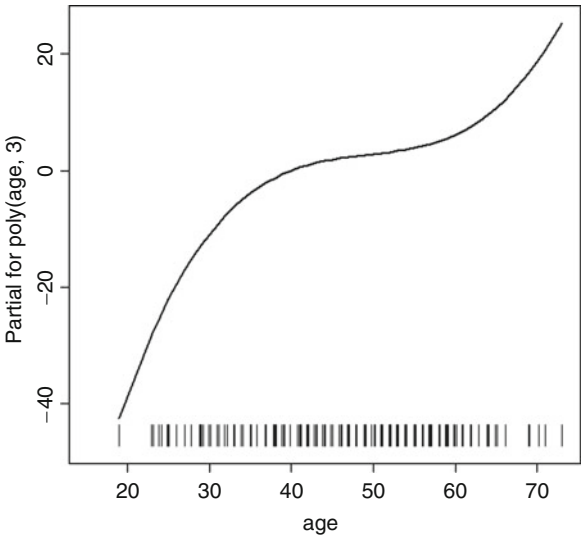


**Fig. 15.5** Term plot for model with polynomial *age*

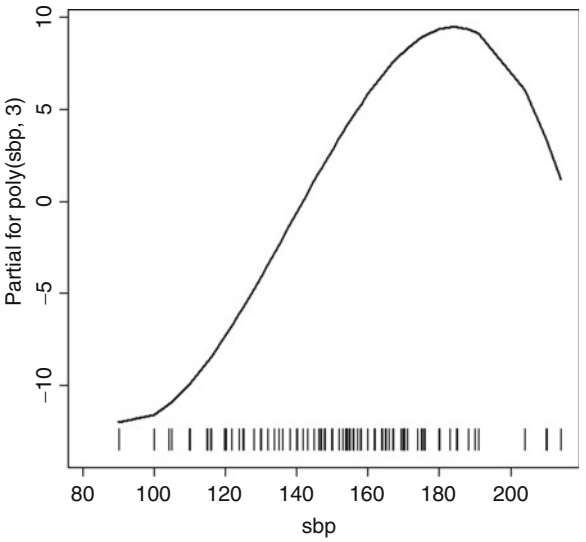


**Fig. 15.6** Term plot for model with polynomial *sbp*

Note that for the model fitted with higher-order terms has two fewer parameters than the model for which the continuous covariates have been discretised. It is not only more parsimonious, but has an adjusted $R^2 = 0.65$, up from 0.63.

## 15.6 Splines

The complexity of the relationship between the continuous covariate and the modelled outcome may be efficiently represented using splines. These are low-order polynomials that are fitted locally but joined at knots smoothly, meaning that at the knots the function represented by the spline, and perhaps also some of its derivatives, are continuous. There are also advantages of numerical stability. The term spline derives from thin strips of flexible wood that have been used in construction to represent complex smooth curves. Fitting splines to covariates can be thought of as taking a nonparametric approach.

In the few situations where a small extrapolation might be considered, splines can often provide less extreme behaviour immediately beyond the range of the covariate. Note that this was a concern in the example above, where the model with higher-order terms predicted very low *sna* for the youngest subjects of the study. Similarly the sharp decline of *sna* with increasing *sbp* above 200 mm Hg provides a further reason to reconsider the model that was fitted. Runge's phenomenon, Runge (1901), which occurs with higher-order polynomials can become problematic. A very nice overview of splines together with a discussion is provided by Eilers and Marx (1996).

There are many ways to specify a basis for a spline fit, Wahba (1990), some examples are B-splines de Boor (1978), P-splines Eilers and Marx (1996), natural cubic splines, and O'Sullivan splines O'Sullivan et al. (1986). The order of the spline approximation must be chosen, as must the number and the location of knots. Penalised splines can be employed, see Sect. 15.8, and then further parameters are involved: the smoothness parameter and the derivative to be smoothed. Smoothing is not considered in this introductory section, but deferred until Sect. 15.8. Knots are often evenly spaced, or placed at certain percentiles of the covariate.

## 15.7 Sympathetic Nerve Activity: Splines

To illustrate the use of splines, natural cubic splines are selected. A single internal interpolation point is chosen as the median (50th percentile) for each of the two covariates. The end points of the range of a covariate are automatically used as knots, and without internal knots the spline degenerates to a polynomial fit. The formula for use with R is `sna ~ as.factor(sex) + ns(age,knots = median(age)) + ns(sbp,knots = median(sbp))`.

In tabulated form, the results of the fit are provided in Table 15.4. It is noted that the fit is not so satisfactory, with the adjusted $R^2 = 0.62$. The effects of the covariates are given graphically in Figs. 15.7 and 15.8.

**Table 15.4** Table of coefficients for model with spline fits for covariates

| Coefficient | Estimate | 95% CI | *p*-value |
|---|---|---|---|
| Intercept | 7.6 | (−0.6,15.8) | 0.0689 |
| Male | 6.2 | (2.8,9.7) | 0.000480 |
| ns(age,knots = median(age)) 1 | 57.1 | (41.5,72.6) | 1.63e-11 |
| ns(age,knots = median(age)) 2 | 20.7 | (11.2,30.2) | 2.80e-5 |
| ns(age,knots = median(sbp)) 1 | 46.0 | (29.0,63.1) | 3.33e-7 |
| ns(age,knots = median(sbp)) 2 | 20.7 | (10.4,31.0) | 0.000109 |

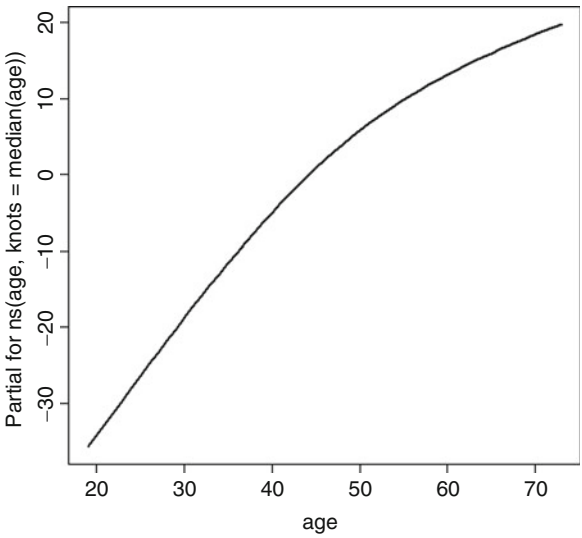


**Fig. 15.7** Term plot for model with natural spline fit for *age*

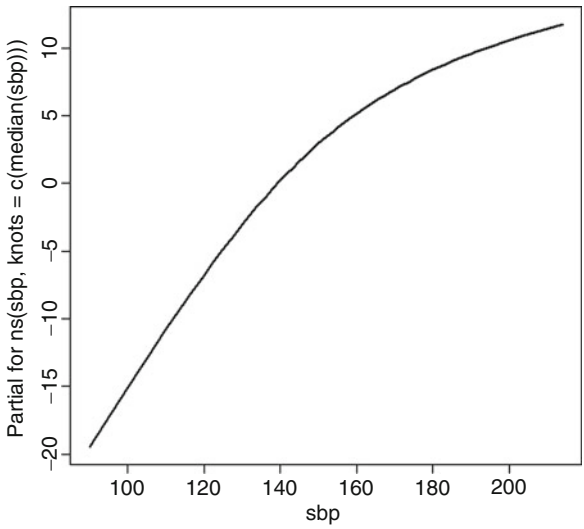


**Fig. 15.8** Term plot for model with natural spline fit for *sbp*

By comparison of Figs. 15.7 and 15.8 with preceding ones, it can be seen that this particular spline fit gives rather different results for the effects of *age* and *sbp* than the other models considered. The fit is better than that of the basic model, but it is clear that there are challenges in finding the best spline representation. Those providing libraries for GAMs have also provided tools to make spline selection much easier and much more efficient: see Sect. 15.8.

## 15.8 Generalised Additive Models

Generalised additive models have continued to receive attention since their introduction by Hastie and Tibshirani, see Hastie and Tibshirani (1986, 1990). Additive models are ones where the effects of each covariate are added: there are no interaction terms and so the additivity of effects is assumed. This chapter focusses on nonlinearities whilst Chap. 16 enables the exploration of interactions. Hence here the initial attention has been to the representation of the effect of each covariate with a graphical representation of that effect to enable interpretation. GAMs continue this theme. The generalised term simply refers to the fact that the methodology of additive models (spline fits to covariates) can be just as easily applied to generalised models, such as logistic regression, as well as it can be applied to linear regression.

Given the large number of parameters that need to be selected for a spline fit, tools to provide automated choices save considerable effort and can provide some objectivity. The principle of parsimony where a model with fewer parameters is preferred to a more complex model is often to the forefront of automated procedures. A statistical epidemiologist will be concerned with estimating the effects of each covariate rather than intricate and subtle choices of parameters in spline fitting and will want to utilise developed software tools with automated choices rather than lavish time and resources on a general spline fit. There is software available to fit GAMs in several statistical packages but here attention is restricted to three libraries that are available in R and which provide more than enough material for discussion in a single book chapter.

## 15.9 Smoothed Low-Order Splines

The fitting of very low-order splines as an initial data-exploration technique is well established and often referred to as 'lowess' or 'loess'. This approach is available for covariates in generalised linear models and has been provided by Trevor Hastie in the function `R::gam::gam`. The default settings are for a spline with degree $= 1$ and span $= 0.5$ so that fitting is performed with a proportion of the data (span) equal to 0.5. Data points receive a tri-cubic weighting proportional to their distance from the estimation point. It is possible to change the degree to be in $\{0, 1, 2\}$ and the span to be in (0,1]. The best policy might be to accept the default settings unless there is evidence not to do so and focus attention on interpreting the effects of covariates.

With the same function in the `R::gam` library it is possible to fit penalised splines (smooths). The target number of degrees of freedom needs to be specified. Rather than expand this aspect here, smooths are considered with the `R::mgcv` package discussed in Sect. 15.11.

## 15.10 Sympathetic Nerve Activity: Loess Splines

The model was reformulated to include loess representation of the two continuous covariates through the formula `sna ~ as.factor(sex) + lo(age) + lo(sbp)`. The fit is excellent with the adjusted $R^2 = 0.66$ and the significance of terms is given in Table 15.5 with the nonlinear nonparametric effects of the covariates shown in Figs. 15.9 and 15.10. Note that there is a facility to display the partial deviance residuals, which was exploited and that upper and lower point-wise twice-standard-error curves were included.

The fits obtained by `R::gam::gam` provide good material for an epidemiologist to consider. The main features of the fits should be explained. Smaller details that lead to a little jaggedness might be ignored in many cases. This approach to interpretation suggests that a smoother fit might be warranted.

**Table 15.5** Table of coefficients for model with loess fits for covariates

| Coefficient | Npar DF | Npar F | *p*-value |
|---|---|---|---|
| lo(age) | 2.5 | 7.71 | 0.000232 |
| lo(sbp) | 3.1 | 6.88 | 0.000187 |

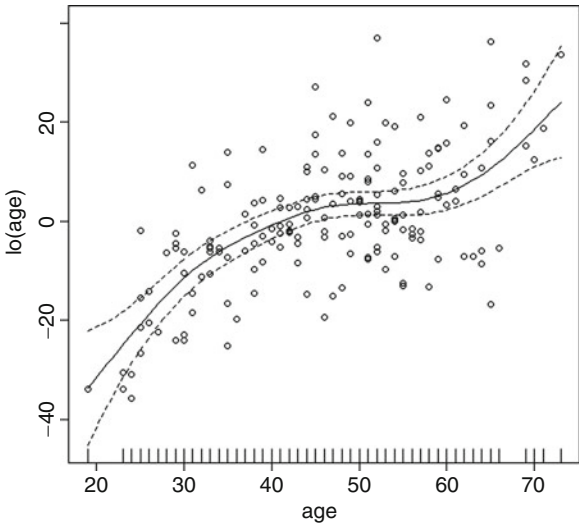


**Fig. 15.9** Term plot for model with loess spline fit for *sbp*

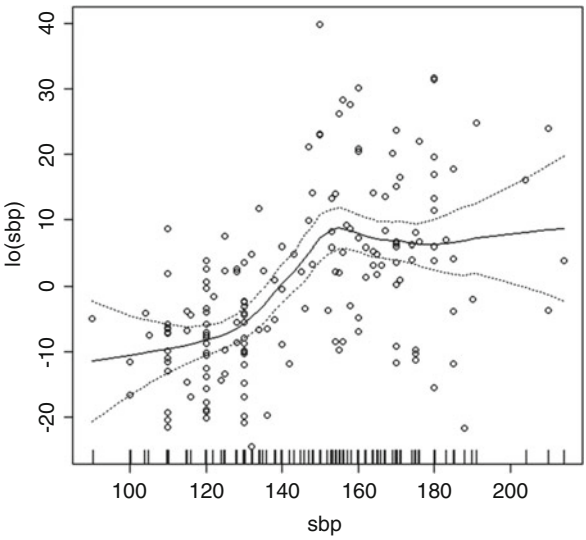


**Fig. 15.10** Term plot for model with loess spline fit for *sbp*

## 15.11 Generalised Cross Validation

The advantages of automatic determination of parameters have been emphasised. Simon Wood (2006) has published a most useful library for automatically fitting GAMs with smooths for covariates, namely `R::mgcv`. Note that this library has a function `R::mgcv::gam` so that it is important to ensure that the correct library has been loaded into R.

A generalised linear model can be fitted by `R::mgcv::gam` identifying which covariates a smooth is to be used: see example below. By cross validation, the 'best' smoothing parameter is chosen, yielding a totally automated procedure. In fact the procedure used is generalised cross validation, which is numerically efficient and yields results close to those of cross validation. Hence a powerful tool is made available to explore smooth non-parametric nonlinearities in covariates for generalised linear models.

## 15.12 Sympathetic Nerve Activity: Cross Validation

The formula needed to indicate smooths for *age* and *sbp* that is used in `R::mgcv::gam` is `sna ~ as.factor(sex) + s(age) + s(sbp)` which reports significance of smooths as is Table 15.6. The effects are shown graphically in Figs. 15.11 and 15.12. Note the great similarity to the results with loess smoothing, although of course the representation of each covariate effect is much smoother, and perhaps therefore more credible in some circumstances. Partial residuals are shown, as are 'twice standard error' curves.

**Table 15.6** Table of coefficients for model with smooths as covariates

| Coefficient | edf | Ref. df | *F* | *p*-value |
|---|---|---|---|---|
| s(age) | 4.289 | 4.789 | 16.41 | 1.05e-12 |
| s(sbp) | 4.812 | 5.312 | 10.39 | 5.40e-9 |

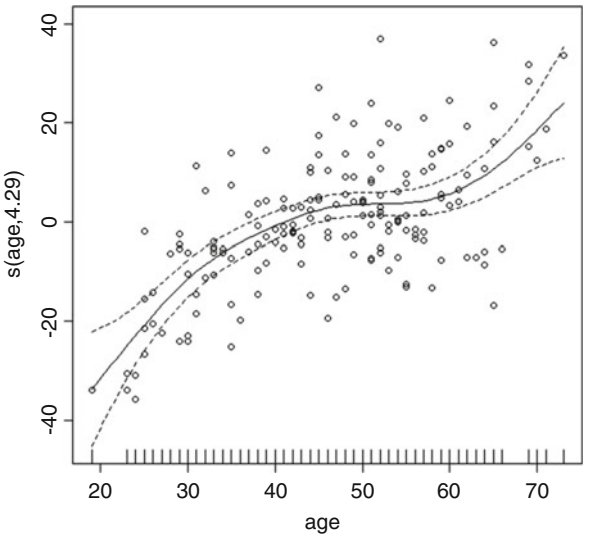


**Fig. 15.11** Term plot for model with smooth for *age*

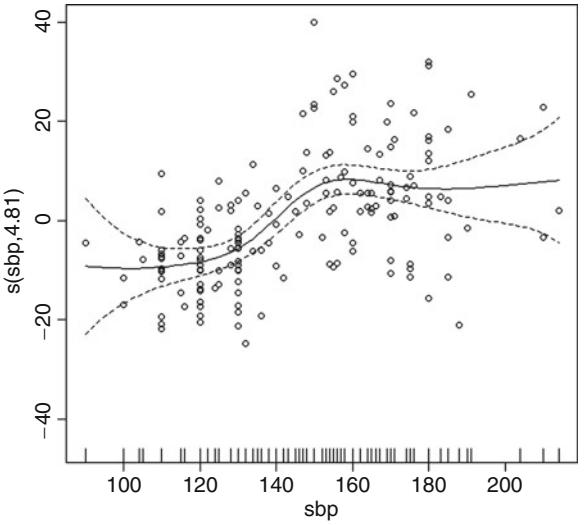


**Fig. 15.12** Term plot for model with smooth for *sbp*

Although the main philosophy of GAMs is to assume additivity of covariate effects, modelling can be extended in dimension by fitting higher-dimensional splines to groups of covariates. This enables interactions to be visualised and compared to strictly additive models. For example two covariates might be suspected of interacting and it would then be appropriate to fit a two-dimension spline. The function `R::mgcv::gam` enables higher-dimensional splines.

## 15.13 Sympathetic Nerve Activity: Two-Dimensional GAM

The fit with a two-dimensional spline for *age* and *sbp* give the best fit to date with adjusted $R^2 = 0.69$ (Table 15.7). Thus there is evidence of an interaction between *age* and *sbp*, see Chap. 16 where this interaction effect is considered further.

Figure 15.13 shows that there are no younger participants with hypertension (high values of *sbp*) and no older participants with *sbp* in the normal range. This might have been a property of the recruiting strategy, or it may be that older people who volunteer for studies tend to have higher systolic blood pressure. The study is cross-sectional rather than longitudinal but there are longitudinal explanations that account for the relationship. Sympathetic nerve activity tends to increase with age and is higher for hypertensives. For younger participants with higher *sbp*, the increase of *sna* with age is more rapid (contours closer together).

**Table 15.7** Table of coefficients for model with 2d smooth

| Coefficient | edf | Ref. df | *F* | *p*-value |
|---|---|---|---|---|
| s(age,sbp) | 12.82 | 13.32 | 26.15 | <2e-16 |

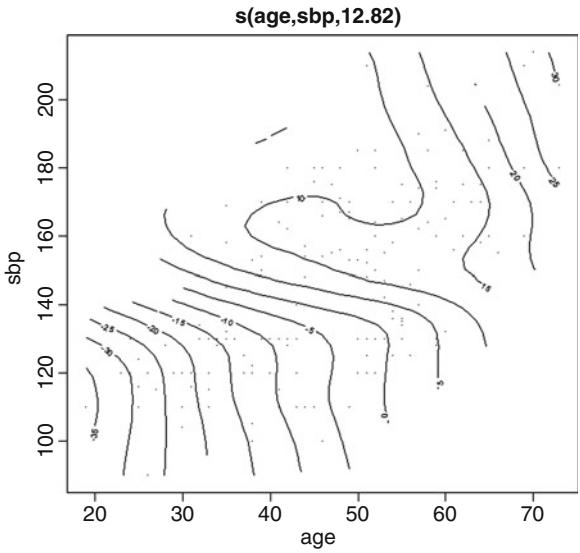


**Fig. 15.13** Plot for model with 2d smooth of *age* and *sbp*

Note that standard error curves were omitted: the plot is already complex and needs full-colour treatment if further information is to be included. For the 2d plot, the standard se curves are $\pm 1$ standard error rather than $\pm 2$ standard errors as with the one-dimensional curves.

## 15.14 Further Aspects of GAMS

This chapter provides an introduction only to GAMs motivating their use through exploration of nonlinearities in covariate effects. Here is a brief mention of further aspects.

A third library is available in R, namely Vector Generalised and Additive Models, see Yee and Wild (1996). `R::VGAM`, that has been made available by Thomas Yee and makes use of B-splines and O'Sullivan splines that have certain advantages. The VGAM library is huge and there is a focus on multivariate outcomes for generalised linear models and generalised additive models.

Random effects can be included in GAMs through the function `R::mgcv::gamm`. Thus GAMs can be used in a multilevel context.

## 15.15 More on the Case Study

Further description of the case study of sympathetic nerve activity was delayed until this point as the primary interest was the methodology for exploring nonlinearities in covariates. Exploring different models however often helps to develop understanding of a situation, indeed that is one of the aims of modelling.

From each of the models it is clear that both *age* and *sbp* make significant contributions to *sna*, explaining well over 50% of the variation in results. Exploring residuals revealed nothing unusual so that for this application there was no indication that a linear model was unsuitable as regards the distribution of residuals. Discretisation of covariates provided little extra information other than indicating that the effect of *sbp* was far from linear. It is possible that a different discretisation would have produced different results: model fitting has challenges. Fitting higher-order terms was found to be no easier. By contrast the procedures for fitting GAMs made modelling far simpler.

Figures 15.10 and 15.12 show partial residuals. These again indicate that the distribution of residuals satisfy distributional assumptions of normality and homogeneity of variance. It is revealed also that there may be some digit preference for some of the participants: those with *sbp* values of 110, 120 and 130, possibly 100, whereas for other values there is no evidence of digit preference. Possibly a different sphygmomanometer was employed for these participants, at a time when younger normotensive volunteers were recruited to the study.

Fitting of statistical models cannot of course reveal biological mechanisms, but knowledge of biological mechanisms may aid in the interpretation of the statistical models. For example, a plausible biological mechanism is that a state of hypertension where *sbp* is constantly raised can result in thickening of the left ventricle and in central sympathetic nerve activity. This suggests that a step increase in *sna* is plausible for patients with *sbp* above the acknowledged threshold for hypertension of 140 mm Hg.

In Sect. 15.13 it was mentioned that the study was cross-sectional but the most plausible interpretation was longitudinal. This suggests that a longitudinal study on sympathetic nerve activity would be of interest. If *sna* is an indicator of progression of cardiovascular disease, then a longitudinal study recording *sna* and cardiovascular events is suggested with analysis using random-effect GAMs.

## 15.16 Chapter Summary

A range of methods to explore nonlinearity in covariates has been outlined and demonstrated with an example. There are considerable modelling challenges posed when there are so many modelling options, and automated procedures were advocated. Different approaches to modelling with GAMs were mentioned. In particular, loess fits can be exploited through `R::gam::gam` and smooths can be automatically selected through `R::mgcv::gam`. Both of these approaches with GAMs have been shown to be capable of producing good modelling results with the effects of covariates made clear through graphical plots.

## References

Burns, J., Sivananthan, M. U., Ball, S. G., Mackintosh, A. F., Mary, D. A., & Greenwood, J. P. (2007). Relationship between central sympathetic drive and magnetic resonance imaging-determined left ventricular mass in essential hypertension. *Circulation, 115*, 1999–2005.

de Boor, C. (1978). *A practical guide to splines*. New York: Springer.

Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science, 11*, 89–121.

Graunt, J. (1662). *Natural and political observations on the bills of mortality*. London.

Hastie, T. J., & Tibshirani, R. J. (1986). Generalised additive models (with discussion). *Statistical Science, 1*, 295–318.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalised additive models*. Boca Raton: Chapman and Hall/CRC.

Kennedy, W. J., & Gentle, J. E. (1980). *Statistical computing*. New York.

Miller, A. J. (2002). *Subset selection in regression* (2nd ed.). Boca Raton: Chapman and Hall/CRC.

O'Sullivan, F., Yandell, B., & Raynor, W. (1986). Automatic smoothing of regression functions in generalised linear models. *Journal of the American Statistical Association, 18*, 96–103.

R Development Core Team. (2010). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Runge, C. (1901). Uber empirische funktionen und die interpolation zwischen aquidistanten ordinaten. *Zeitschrift fur Mathematik und Physik, 46*, 224–243.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). NewYork: Springer Science and Business Media.

Wahba, G. (1990). *Spline models for observational data*. Philadelphia: SIAM.

Wilkinson, G., & Rogers, C. (1973). Symbolic description of factorial models for the analysis of variance. *Applied Statistics, 22*, 329–399.

Wood, S. N. (2006). *Generalised additive models: An introduction with R*. Boca Raton: Chapman and Hall/CRC.

Yee, T. W., & Wild, C. J. (1996). Vector generalised additive models. *Journal of the Royal Statistical Society, Series B, Methodological, 58*, 481–493.